ELSEVIER

Artificial Intelligence 80 (1996) 99–117

Artificial Intelligence

Book Review

Embedded or embodied? a review of Hubert Dreyfus' *What Computers Still Can't Do*[☆]

H.M. Collins*

Centre for the Study of Knowledge, Expertise and Science (KES), University of Southampton, Southampton SO17 1BJ, UK

Received February 1995

Part I

1. Introduction: three faults

What Computers Can't Do is a classic. First published (with the shorter title), in 1972, it is one of very few works of philosophy that have successfully challenged the empirical claims of a modern science. The book made predictions about the potential successes and failures of intelligent machines which were thought outrageous at the time but which have turned out to be almost exactly correct. Perhaps the greatest compliment one can pay the book is to say that the predictions seem unremarkable nowadays because they are part of the common-sense of a large part of the community which thinks about these things.

What Computers Can't Do is based on the ideas of the later Wittgenstein and of phenomenological philosophy. When I read the book (in 1983), it seemed enviably comprehensive and decisive. At the time I was applying Wittgensteinian ideas to the analysis of science and Dreyfus seemed to leave little to say on the subject of artificial intelligence. Inevitably, the major themes of subsequent analyses of machine intelligence, such as those by Winograd and Flores [23], Suchman [20], and myself [3], were anticipated by Dreyfus. Even his own more recent book about expert systems [9] is unimpressive in comparison. Fortunately for later authors, there are some cracks in the argument of *What Computers Can't Do*. I want to show that these cracks are the surface manifestations of deeper lying faults that have taken nearly 20 years to become evident.

I am going to argue that Dreyfus makes a "professional mistake", a "philosophical elision", and a "sociological error". The *professional mistake* is

[☆] (MIT Press, Cambridge, MA, 1992); liii + 354 pages, $13.95.

* Corresponding author. E-mail: h.m.collins@soton.ac.uk.

SSDI 0004-3702(95)00083-6

selective application of the philosophical critique. In particular, he does not notice that the Wittgensteinian analysis applies just as much to science, technology and mathematics as it does to other areas of culture. This allows him to set up a false dichotomy between "formal" domains of knowledge where computers can do things and "informal" domains where they can't. I will refer to the spurious division between the two as the "knowledge barrier". The *philosophical elision* is found within the treatment of embodiment. It is a matter of mixing up the properties of individuals with the properties of the social groups to which they belong. While the physical embodiment of entities may give rise to their collective "form of life" (to use a Wittgensteinian phrase which I will explain below), this does not mean that every individual that is to share that form of life must have the same type of physical body. Immobile cuboid boxes in general may never develop human-like ways of being; but this argument is not enough to show that an *individual* immobile cuboid box could not. The elision leads Dreyfus to over-emphasise questions to do with the detailed architecture and physical form of computers and underemphasise the problem of what it is to be a member of a social collectivity. The *sociological error* is not seeing that many of the apparent capabilities of computers are a result of their real-time interactions with their users. This means that even if computers are not proper members of social collectivities, they can seem to act like humans with the (usually unnoticed) help of their users. Often, the results are excellent. Thus, some philosophically and psychologically uninteresting computers can do sociologically interesting and useful things in unanticipated ways.

The faults, to extend the earthquake metaphor, have not given rise to damage because the book stands on the foundation of its nearly correct predictions about what computers would be able to achieve in various domains. The predictions, however, have stood the test of time for reasons more complex than Dreyfus imagined. This does not diminish the book's status: almost all the important arguments concerning what computers *cannot* do are there; almost all of them are correct if sometimes incorrectly focussed; the arguments are forcefully set out and exemplified; and nearly all the head-on rebuttals of the book, encouraged by its polemical tone, are just as misdirected as they every were. The mistakes are not to do with what computers can't do, but what they *can* do and how they do it.

2. What computers can and can't do: the professional mistake

If one examines the book carefully one finds very little about what computers *can* do. The criticisms dominate: so long as computers don't have bodies they won't be able to do what we do; so long as computers represent the world in discrete lumps they won't be able to respond to the world in our way; we are always *in* a situation, whereas a computer only 'knows' its 'situation' from a set of necessary and sufficient features; we follow rules without being subject to the Wittgensteinian regress of rules that are required to explain how rules should be applied and further rules to explain how those rules are to be applied and so

forth. The trouble is that if you take all of this on board, computers ought not to be able to do anything.

So what, according to Dreyfus, can computers do in spite of these disabilities? On pages 291–293 we are told some of the answers: they can reproduce "elementary associationistic behaviour where meaning and context are irrelevant to the activity concerned" (examples include mechanical dictionary translation and pattern recognition by template-matching); they can work well in areas that involve "the conceptual rather than the perceptual world" ... [where] ... "problems are completely formalized and completely calculable". In the Introduction to the, 1979, second edition of the book there is more: Dreyfus admits his earlier neglect of certain programs which "while solving hard technical problems and producing programs that compete with human experts, achieve success precisely because they are restricted to a narrow domain of facts, and thus exemplify what Edward Feigenbaum, the head of the DENDRAL project, has called 'knowledge engineering'."

The odd thing about all of this is that Dreyfus, as a follower of Wittgenstein, ought not to think of the conceptual world as much different from the perceptual world. The conceptual world too is "situated". The later philosophy of Wittgenstein shows that what we take as logically and scientifically compelling is what we do not know how to doubt [1, 21, 22]. What we take as certain is what follows for us, as a matter of course, in the way we live in the world. In the last resort there is no more compelling proof. Even logical syllogisms cannot be proved if we are unwilling just to "see" and act as though they follow in the situations in which we find ourselves. Thinking and acting in the world are but two sides of the same coin.

The term that is used to describe the combinations of thinking and acting which constitute our experience is "form-of-life". One way to approach this difficult idea is to start from discussions of science. Thomas Kuhn's "paradigms" can be thought of as Wittgensteinian forms of life within the culture of science (though Kuhn himself did not make anything of the parallel). A paradigm gives us our ways of thinking and acting scientifically; it is a whole scientific way of being [13]. Thus within one paradigm something may be absolutely fixed while in another it is not (e.g., mass is fixed for Newton but not for Einstein). Likewise, what may count as a scientific finding is determined by the "conceptual structure" associated with what counts as doing a proper piece of scientific work (e.g. any experiment that did not conserve mass would, by that fact, be flawed under the Newtonian paradigm but this is no longer the case under relativity). There is no way of proving something either conceptual or experimental outside of the frame of a paradigm/form of life. The later philosophy of Wittgenstein makes the world of ideas a subject for sociological study while making the social world a subject for conceptual analysis.

In recent years the Wittgensteinian starting point has been used in detailed analyses of mathematics, science and technology, showing us how to see them as just as much social phenomena as, say, natural language. To give some idea of how this can be we may think about the process of replicating an experimental

claim to find out if it is true. As everyone knows who has ever done a practical class, experiments are hard and they usually don't work first time if they work at all. In the normal way this presents no problem because we know roughly what the result of an experiment should look like and we carry on until we get it right. A problem arises, however, when the correct outcome of an experiment is in dispute, as is typical in frontier areas of science. In these areas the skill-laden nature of experimentation becomes much more obvious. Typically scientists whose findings are disputed claim that their rivals did not do the experiments carefully enough and that is why they cannot find the phenomenon. The rivals will claim that there is no phenomenon and the original experiments were done in a sloppy fashion. There is no way to settle such a dispute with further experiments alone for each experiment may or may not have been done sufficiently well and there is no direct way of measuring experimental skill. The usual surrogate measure—Does the experiment produce a result in the right range?—is not available. In sum, we get the "experimenter's regress": to know whether "x" exists one needs to build a good "x-detector", but to know whether you have built a good x-detector you have to see whether it does what it is supposed to do, but to know what it is supposed to do you have to know whether or not it ought to detect x when it is working properly, but to know whether it ought to detect x when it is working properly you need to know whether x exists, but to know whether x exists you need to build a good x-detector, and so forth [2]. The skillful nature of experimentation and the workings out of the experimenter's regress have been shown in, among other things, case studies of laser building, the detection of gravitational radiation, and experiments in parapsychology [2].

The phenomenon applies just as much in the case of famous scientific experiments though this is surprisingly little known. For example, the lay history of physics recounts that, say, Einstein's (1905) special theory of relativity was proved to be correct because the Michelson–Morley experiment of 1887 showed clearly that the speed of light was constant in all directions. What is not widely known is that Michelson himself was never satisfied with the results of his experiment, that according to the protocol that he and his collaborators set out, the experiment was never done properly, and that as late as 1925, the American Association for the Advancement of Science awarded its prize to one of Michelson's erstwhile colleagues—Dayton Miller—for an improved experiment showing that the speed of light varied by 10 km per second according to the direction of measurement in relation to the orbit of the Earth. A very similar story can be told about Eddington's 1919 solar eclipse observations, which are widely taken to prove the general theory. Most other "decisive" passages of scientific, technological and mathematical history turn out to have looked far less clear cut at the time and place they were done [7]. They look certain only at a distance. All this means that the outcome of a passage of scientific work is very much more like a matter of social agreement about what it is proper to believe than the application of an experimental or theoretical algorithm. It is, then, unsurprising that what count as scientific laws seems to vary in different historical and social circumstances.

Given this, the computer modelling of science, as well as the achievements of DENDRAL, MYCIN and the other products of "knowledge engineering" ought to be just as puzzling to a Wittgensteinian as natural language translation, for all are matters of understanding the subtleties of social situations. Where Dreyfus says (pp. 200–201) "When one uses the laws of physics to guide missiles, for example, the present performance of the missile is an instantiation of timeless, universal laws which make no reference to the situation except in terms of such laws." he is misunderstanding the nature of scientific laws and expecting far too much of missiles. Missile guidance is not a matter of timeless laws but of the application of technological skill, and the extent to which missiles are accurate is a matter of the extent to which designers, builders and operators are skilled in their respective crafts [16]. But even what counts as a proper measure of accuracy, what counts as a proper determination within an agreed measure and what counts as a piece of data are matters of social agreement. In recent times this was made graphically evident by the argument over the performance of "Patriot" during the Gulf War [18, 19]. Physical laws are just the last things we agree to hold constant while we argue about everything else. In serious disputes, even these are given up [2]. Though we usually think of them as universal, the perceived scope of physical laws is a matter of how well our scientific form of life is going at the time. Dreyfus, in spite of his philosophical roots, seems to have invested too deeply in a pre-Wittgensteinian, steady state, view of science which allows him to think of the language of scientific laws as outside of social life in a way that, say, natural language is not. This leads him to give *too much* credence to the power of computers in those areas which he allows himself to think of as "formal". This is his "professional" mistake [4, 8].

3. The problem of embodiment: the philosophical elision

All three of the deep faults in *What Computers Can't Do* come back in one way or another to neglect of the social embeddedness of both humans and computers. One of the clearest manifestations is to be found in Dreyfus' analysis of the problem of embodiment. Here we find the philosophical elision: mixing up the properties of individuals and social collectivities.

Dreyfus, once more following Wittgenstein, shows that the way our bodies fit into the world makes the world available to us. For example, this is what he has to say about the problem of recognizing chairs:

> What makes an object a *chair* is its function, and what makes possible its role as equipment for sitting is its place in a total practical context. This presupposes certain facts about human beings (fatigue, the way the body bends), and a network of other culturally determined equipment (tables, floors, lamps), and skills (eating, writing, going to conferences, giving lectures, etc.). (p. 237)

He goes on to say:

> Since it turns out that pattern recognition is a bodily skill basic to all intelligent behavior, the question of whether artificial intelligence is possible boils down to the question of whether there can be an artificial embodied agent. (p. 250)

In this, 1972, treatment, the idea of embodiment is very much tied up with the physical constitution of the body and its interactions with the familiar furniture of the world. In the 1992 edition, Dreyfus seems to have changed his view, embodiment becoming a more conceptual notion. On pages xx to xxi of the Introduction to the 1992 edition Lenat is quoted, putting an argument against Dreyfus' earlier view: Wheelchair-bound "Madeleine", it seems, was blind from birth, could not use her hands to read braille, and yet acquired commonsense knowledge from books that were read to her. Lenat argues that this shows that the body is not as important to pattern recognition as Dreyfus claimed. Dreyfus' counter-argument is that Madeleine has a body with an inside and an outside, which can be moved around and, in addition, Madeleine has imagination; she can empathise with those who have more complete bodies. But under this argument a body is not so much a physical thing as a conceptual structure. If you can have a body as unlike the norm and as unable to use tools, chairs, blind persons' canes and so forth as Madeleine's, yet you can still gain commonsense knowledge, then something like today's computers—fixed metal boxes—might also acquire commonsense given the right programming. It is no longer necessary for machines to move around in the world like robots in order to be aware of their situation and exhibit "intelligence".

The resolution of the argument between Lenat and Dreyfus is to be gained, once more, by seeing how ideas are embedded in the social world. Wittgenstein said that if a lion could speak we would not understand it. The reason we would not understand it is that the world of a talking lion—its "form of life"—would be different from ours. Bringing back Dreyfus' chair example, lions would not have chairs in their language in the way we do because lions' knees do not bend as ours do, nor do lions "write, go to conferences or give lectures". Circus lions talking among themselves would, presumably, group what we call a household chair along with the other weapons they encounter in the hands of "lion tamers", not with objects to do with relaxation. They would not distinguish between sticks and chairs and this is why their language would be incomprehensible to us. But this does not mean that every entity that can recognise a chair has to be able to sit on one. That confuses the capabilities of an individual with the form of life of the social group in which that individual is embedded. Entities that can recognise chairs have only *to share the form of life* of those who can sit down. We would not understand what a talking lion said to us, not because it had a lion-like body, but because the large majority of its friends and acquaintances had lion-like bodies and lion-like interests. In principle, if one could find a lion cub that had the potential to have conversions, one could bring it up in human society to speak about chairs as we do in spite of its funny legs. It would learn to recognise chairs as it learned to speak *our* language. This is how the Madeleine case is to be understood; Madeleine has undergone linguistic socialization.

In sum, the shape of the bodies of the members of a social collectivity and the situations in which they find themselves give rise to their form of life. Collectivities whose members have different bodies and encounter different situations develop different forms of life. But given the capacity for linguistic socialisation, an individual can come to share a form of life without having a body or the experience of physical situations which correspond to that form of life. What we don't know is how to make something with the capacity to be socialized in this way.

What we do know is the following: First, dogs, say, do not have the capacity to be sufficiently linguistically socialised to pass even the simplest Turing test in spite of their ability to move around in our world encountering and coping with the same physical environment, and in spite of the fact that their brains are much more like our brains than they are like computers, and in spite of the fact that they can be trained to do highly complex specifiable tasks. A dog does not have the capacity of the hypothetical lion with conversational potential discussed in the last paragraph. And this is a matter of conceptual mismatch not just vocalisation. For example, dogs can't be trained to tidy up a house even though they would be physically capable of doing so and the rewards for successful trainers (and their dogs), would be great. The problem is that the concept of an acceptably and appropriately "tidy" room is bound to social context: it is different for a student and a diplomat; Sunday morning is not the same as Saturday evening; today's newspaper is not the same as yesterday's newspaper while even yesterday's newspaper can gain value in the hands of an artist or an antique dealer. To understand "tidy" one has to understand all this and much more.

Second, such evidence as we have from those who have been isolated from normal human society suggests that even persons who are the same as us in terms of brain and bodily structure may not have the capacity to be socialised if the socialisation does not start early enough in their lives. We know, then, that immersion in human-like physical and social situations is not *sufficient* to produce socialisation even where the brain and body are like or even identical to those of humans.

Third, humans, such as Madeleine, whose bodies are very different from the norm and who cannot move around in the way that dogs and inadequately socialised humans can, do have the capacity to be linguistically socialised and come to share our conceptual structure and way of seeing the world. Madeleine would know the difference between a tidy and an untidy room even though she could not tidy it up. We know, then, that human-like bodies are not *necessary* for human-like socialisation.

Putting all this together, we can say with confidence that if we can't train a computer without a body to act like a socialised human, giving it the ability to move around in the world encountering the same physical situations is not going to solve the problem. On the other hand, if we can find out what is involved in the sort of socialising process undergone by a Madeleine—let us call it "socialisability"—we may be able to apply it to an immobile box.

The question that remains is whether socialisability and immersion in a human

form of life, even where this consists solely of encounters with linguistic situations, is necessary to create a shared conceptual structure or whether such a structure can be engineered some other way. Computer optimists might point out that a few silicon chips with no social experience whatsoever can be programmed with what can sometimes pass for linguistic competence. Is there hope, then, for steady incremental progress toward social competence without socialisation? In spite of the small successes, and in spite of the philosophical elisions, I believe Dreyfus is right to dismiss these hopes. The commonsense knowledge we need for full language use is not the sort of collection of facts and rules that, say, Lenat is assembling with his Cyc project; rather, it does have to do with being a member of the relevant social collectivity [15]. The linguistic capacity of computers—even, I predict, those which take advantage of the products of the Cyc project—is not nor will be equivalent to linguistic socialisation. Many of Dreyfus' arguments (like my "tidiness' example and the studies of science referred to in the previous section), show why this is so. At the end of this review I will suggest a simple way of revealing the difference between the minimal linguistic competence currently and foreseeably attainable by computers and proper linguistic socialization.

It is worth noting, finally, that the Dreyfusian embodiment critique still applies where claims are made on behalf of collections of computers. There are those who say in response to the criticism that computers aren't social or socialisable, "I have a whole network of interlinked computers!" Others are trying to model society by programming the interaction of "actors" using quasi-economic reasoning. These ideas are misplaced for reasons which might have been perfectly expressed by Dreyfus if the elision between the embedded and embodied individual had been resolved. I would rewrite his passage on page 250 as follows: "Since it turns out that pattern recognition is a bodily skill basic to all intelligent behavior, the equation of whether an artificial *society* like ours is possible boils down to the question of whether there can be artificially embodied agents like us." What the project of AI still lacks is a computer or a network of computers, embodied or unembodied, that has been or could be socialized into our form of life, linguistically or otherwise. Incidentally, the same applied to ants and bees, those most "social" of creatures.

4. Repairs, domains and micro-worlds: the sociological error

While Dreyfus' treatment of the nature of science exhibits a professional mistake and his treatment of embodiment contains a philosophical elision, the overall treatment of the abilities of computers reveals a sociological error. He does not see or draw the appropriate conclusions from the fact that much of the ability of computers lies not inside the case but in the way we interact with them when we use them; we continually "repair" the deficiencies of computers (and animals and other humans). A good way of getting hold of this is to think about Weizenbaum's famous ELIZA in its DOCTOR instantiation. DOCTOR, as a computer program, or "electronic brain", is a toy, yet as a psychotherapist it is

good enough to have given rise to arguments about its ability to replace humans. Simply to make the obvious response about the nature of psychotherapy is to miss the point; the point is that if the user is putting enough into the social interaction, the computer does not have to do very much. In the case of DOCTOR, the patient does most of the work. The *locus classicus* for this argument is Garfinkel's "counsellor" experiment, in which students imputed counselling skills to a list of "yes–no" responses driven by a table of random numbers [11]. This is also what makes it possible for "con men" to operate; they make sure that the "mark" *wants* to believe the stories they tell [17].

One can see how this works in the context of computers by choosing a very simple but very familiar example, a pocket calculator. We think of pocket calculators as doing arithmetic better than we do, but in some respects they do not. When you use a pocket calculator you have to prepare the world for it: you have to work out in what order to insert the symbols, depending on what type of calculator you have, (try "12x − 4"); you have to work out what the symbols mean (was the "x" in the last sum an algebraic symbol or the multiplication symbol); you have to "repair" the calculator's mistakes (try $7/11 \times 11$ on your cheap pocket calculator or on your powerful mainframe working at maximum precision); you have to approximate appropriately on behalf of the calculator—approximation being the basis of nearly all science and much arithmetic [14]; you have to read the keys and the display, know how to know that you have made a keying error, and know how to know when the sum is wrong and, of course, before you can start, you have to know how to put your question into arithmetical form [3]. Only a little bit of work is done by the calculator. All this is what we ought to expect for, if we apply the ideas of Wittgenstein assiduously, it is puzzling that a calculator—obviously an unsocialized entity—can do anything at all of the *social* practice of arithmetic.

There are, contrary to Dreyfus, no "domains" in which computers work. They work a bit in all domains and they don't work entirely in any domain. The only place where computers work without any repairs is in "micro-worlds", but these are *not* the same as domains.

From page 4 onward in the Introduction to the 1979 edition, Dreyfus provides a superb critique of micro-worlds (a critique later accepted by Winograd). It is summed up on pages 13 and 14:

> . . . one cannot equate as [Winograd] does, a program which deals with "a tiny bit of the world," with a program which deals with a "mini-world." . . . sub-worlds are not related like isolable physical systems to larger systems they *compose*; rather they are local elaborations of a while which they *presuppose*. (Dreyfus' stress).

In other words, so called micro-worlds are not like little bits of the real world, because every little bit of the real world has roots in the world as a whole. This, once more, is the problem of commonsense knowledge. SHRDLU deals with blocks, but if I ask SHRDLU to put the sugar-lump-shaped, sky-coloured block on top of the Toblerone shaped block which is the color of that stuff you get in

the corner of your eyes when you wake up after a heavy night, it will have to know more than is likely to be in its *micro-world* at the time of asking. Nevertheless, the question still belongs to the *domain* of blocks. A great deal of discussion of expert systems in particular and the capability of computers in general runs into trouble because domain-specificity is confused with restriction of operations to a micro-world. And, sometimes, Dreyfus himself mixes them up. The worlds of the conceptual, of spectrum analysis and of science in general are not micro-worlds, they are domains, and they have their foundations in common sense. Turning back to arithmetic for a final illustration: Try "VIII times VII?". What is the answer? It is "56" or is it "LVI" or "$V^2I^2I^2I$". The solution is not to be found in any arithmetical micro-world, though it clearly belongs to the domain of arithmetic.

5. Mimeomorphic actions

So, what *can* computers do? The question is difficult for philosophically-inclined sociologists or sociologically-inclined philosophers. When they learn to become Wittgensteinians (or ethnomethodologists or, perhaps, Heideggerians), they learn to take anything and everything that people think is universal, straightforwardly rule following, or otherwise simple for humans to accomplish, and show how in reality it is local and situated, always requires an infinite regress of rules to explicate fully, and depends therefore on full socialization into a form of life to accomplish. Is there anything that humans do that is not like this?

Let us make a fresh start: "Can machines mimic humans?" The answer is "yes"—just to the extent that humans can mimic machines. And what is it to mimic a machine? The key, once more, is to be found in Dreyfus—though he makes little of it. On page 291 he says: "Area I [the first of Dreyfus' four "areas of knowledge", which I will discuss below] is where the S–R [stimulus–response] psychologists are most at home." In our (social scientists' and socially minded philosophers'), rush to show that everything is situated, socialized, and otherwise complicated, we have forgotten that humans can act in another way—a way that can be modelled with a micro-world and taught after the fashion of the Skinnerians. I refer to this way of acting as mimeomorphic action. (In earlier work I called it "behavior-specific action" [3].)

Let us define an "action" as the sort of thing that I might normally do in some society or another, such as catch a train, score a goal, write a cheque, write a love letter, wink (not blink), greet someone, or wave to someone (actions can be embedded within one another). Let us define a behavior as the bodily movements I use in order to instantiate that action on some particular occasion. Then, put simply, an ordinary action is characterised by the fact that the same *action* can be instantiated by many different *behaviors*. For example, paying money can be done by passing metal or paper tokens, writing a cheque, offering a plastic card and signing it, and so forth, and each of these can be executed in many different ways in terms of the space-time co-ordinate description of my bodily movements. At

the same time, the same piece of behavior may be the instantiation of many different actions. For example, signing one's name, with identical hand movements on each occasion, might be the action of paying money, or it might be agreeing to a divorce, the final flourish of a suicide note, or a making a specimen signature for the bank. There is, then, no straightforward mapping between actions and behaviors. One might train a person, or a pigeon to *behave* in a certain way, but that would not be the same as training them to *act* in a certain way.

Most of the time most of our actions are like this and that is what makes social science and the mechanical reproduction of actions so difficult. When we observe behaviors we do not, thereby, understand actions, and when we reproduce behaviors, we do not, thereby, reproduce actions; yet it is the orderliness of actions not behaviors that makes the world we experience. This is another way of expressing the Wittgensteinian point common to all the philosophical critiques of AI to which I have referred.

This description, however, leaves space for a special class of actions. In "mimeomorphic actions" we *do* attempt to maintain a one-to-one mapping between our actions and observable behaviors. The meaning of certain actions is bound up with their being mimeomorphic (e.g. marching), whereas in others the opposite is the case—they are essentially "polimorphic" (e.g. writing love letters), but many types of action can be executed in either way depending on intention, desired outcome and what seems appropriate in the context. The term "mimeomorphic" refers to the possibility of reproducing such actions by copying previous occurrences; "polimorphic" actions take many shapes (poly) but to get the shape right needs reference to the society—*polis*. (In earlier work I called polimorphic actions "regular actions" [3].)

There is a complication. Since there are always parameters within which every behavior is different from those which preceded it (e.g. the time of day is different, the Sun might be shining on one occasion and not on another, and so forth), and since the level of accuracy of the space–time description of any movement can be increased almost indefinitely, there is always variation of behavior within some frame of reference however hard we are trying to eliminate it. Thus, by "the same behavior" we have to mean the same behavior *within* the outer edges of our indifference. In terms of intention, this may be marked either by our preference for a mimeomorphic action to be carried out in the same way each time so far as we can manage and detect, or our readiness for it to be carried out with an arbitrary degree of behavioral similarity so long as the variation remains within acceptable bounds [6]. The notion of mimeomorphic action, then, allows actions to be unambiguously mapped onto behavior even when there is a degree of variation in the behavior.

In a "digitised" system classes are kept separate because we are tolerant to a degree of variation. To use Haugeland's example, a dollar coin is still worth a dollar even if it is worn or clipped (compare this to a gold bar, which changes value with every change in substance) [3, 12]. One might say, then, that in mimeomorphic actions, action and the corresponding behavior are digitised.

One finds mimeomorphic actions in areas as varied as work on Taylorist production lines, the golf swing, high-board competition diving, ideal bureaucracies and *some* arithmetical operations.

Thus, there are areas of human life where behavior *can* be substituted for action, where humans could be trained to accomplish satisfactory outcomes in the way that pigeons are trained, and where machines that exhibit the appropriate behavior can stand in for human counterparts. In human life there are no domains that are wholly like this, but many where we strive to make them thus. Initial military training is perhaps the best developed large-scale version but the phenomenon is found on a small scale *within* many domains. It is in these bits of domains where computers and other machines can directly replace human action; for here humans are trying to make the aggregate of their own actions into micro-worlds [3]. Since this is difficult, a computerised replacement will often represent an improvement. (Because the boundaries of the micro-world-like bits rarely correspond to the boundaries of the complete role of any human, replacement of whole humans by machines—without reorganizing everything else—is rarely successful.)

To summarise, many domains contain micro-world-like elements, that is, little areas where people are doing their best to accomplish mimeomorphic actions some of the time. Where they succeed, they have built the human version of a micro-world. Many institutions that are taken to be micro-world-like, such as ordinary bureaucracies, ordinary production lines, most of the military, and nearly all of science, are not micro-worlds at all. But all of these institutions contain bits and pieces that *are* micro-worlds—in those little areas, computers fit naturally.

6. The story so far

Dreyfus' position is built on deep faults. His predictions are correct for reasons that are far more complicated than he supposes. His professional mistake and his sociological error combine to lead him to be too generous about the capabilities of computers in narrow domains and too pessimistic about computers' abilities in wider domains. To explain the difference between successful and unsuccessful computers he invokes the idea of "formal" and "informal" domains of knowledge, an idea which is incompatible with his general philosophy.

The alternative that I am putting forward here has three parts. First, computers can do anything where the user supplies most of what needs to be done in the way of expertise. Second, "repairs"—by which is meant the "filling in" of gaps and misunderstandings in computers' performance—are readily available and largely invisible if the required expertise is ubiquitous or already familiar to the users of the machine. (That is, if it is what already counts as commonsense among those in the domain.) Third, computers *can* reproduce what humans do in those areas where we prefer to act in a mimeomorphic way. (Chunks of arithmetic are like this and so are small bits of many domains.) How successful computers are in this

capacity is a matter partly of computer design and partly of history. We change the way we prefer to exercise our skills as history unfolds. If George Orwell's *1984* had come true, and we all spoke a simplified "Newspeak", we might now have successful speech transcribers. This model of what computers can do is much more complicated than that of Dreyfus. I now have to explain how it is that his predictions are so nearly right while his explanation of what computers can do is so over-simplified.

On page 293, Dreyfus offers predictions of computer success and failure in terms of four "Areas". Area I is called "Associationistic" and includes such things as word-for-word translation. Area II, "Simple Formal", includes computable games such as tic-tac-toe and theorem proving. In the terminology of this text, it includes the domain of the "conceptual". Presumably it also includes missile guidance and those bits of science that, according to the 1979 Introduction, are within the purview of knowledge engineering. The paradigm Area III task is chess. Area III is "Complex-Formal" and includes games that are computable in principle but not in practice. Area IV, "Nonformal", includes normal language translation and all the rest of our activities that are context- and situation-dependent.

A crucial part of Dreyfus' position is that Area IV is qualitatively different from the other three; this is the area that is beyond computerisation for fundamental reasons. Area IV lies beyond his knowledge barrier. For Dreyfus, no digital computer will ever accomplish Area IV tasks; this is an in-principle argument. (Though there is some ambivalence about the nature and potential of neural nets—see below.)

I am going to argue, on the other hand, that there is no knowledge barrier, and that there is continuity between all four areas. According to my view, the areas vary only in the ratio of the types of actions we find in them and how these types of actions relate to what is already familiar to us. The discontinuity in my theory is between ordinary and mimeomorphic actions. In my theory no computer that has not been socialised will be able to perform polimorphic actions and no computer that is based on existing designs can be socialised. But elements of both types of action are to be found in each of Dreyfus' four areas (in different and changing proportions), while repairs enable some performances without the benefit of socialisation to pass as competent.

In discussing the four areas in detail I'll dispose of Area III first since, though it has been the subject of most heat, the debate bears on almost nothing of philosophical importance.

Dreyfus is said to have been proved wrong by the progress of chess-playing programs and this accounts for the heat of debate in Area III. Yet the job of a chess computer is to beat a human at chess irrespective of method; this need have nothing to do with mimicking human abilities; it does seem that recent successes have come through brute strength methods. Compare this with, say, language. A successful language using computer is not supposed to beat a human at language; the whole idea is incoherent. If language-using computers were to win in Turing test competitions they would win not by beating humans at language but by being

indistinguishable from humans. The success of chess computers tells us no more about artificial (human-type) intelligence than the success of tractors at tug-of-war tells us about muscles.

Turning back to the other areas we can see how in every case both kinds of actions are involved in the associated domains. As far as Area I is concerned, the treatments of both Dreyfus and myself agree that associationistic, stimulus–response type tasks can be managed by computers. But I say this can happen only after the stimuli have been digitised [3, 12]. Consider word-for-word translation. I have just yelled to my general-purpose daughter "Hey-Lil, wosthe-frarnsezfercow" (she replied, "vash"). At the very least, for even a translation-dedicated, speech-recognising computer I would have to enunciate my words separately and clearly and say "French" not "frarnsez". The computer and I would be fitting my question into a set of unambiguous templates. (I am arguing that barring socialisation of the computer I will always have to do the larger part of this job.) Once something has been digitised it becomes amenable to mimeomorphic action and can be thought about in terms of stimulus–response. (If we go back to human and animal psychology, the same applies. Stimulus–response only works when the organism responds unambiguously to classes of positive and aversive stimuli. We and the organism together do the digitising.)

Area II actions, as explained above, are not as "formal" as Dreyfus takes them to be. The interesting thing about much of Area II is that the digitisation of inputs and repairs of outputs that we do is invisible. To coin a phrase that will be familiar to English readers, in Area II we do "invisible mending"; invisible mending is mistaken for no mending at all. Thus, think of how many different ways there are of playing tic-tac-toe in real life: with paper and pencil, with blackboard and chalk, with sand and sea shells. . . . Each of these depends on our willingness to repair the variations and see the playing arena as a set of symbols; that is how we and our opponent can agree about what we are doing. Playing tic-tac-toe with a computer requires that we present it with an unambiguous input and translate it's output into the familiar game in something of the same way.

When we turn to science the matter is even more revealing. We are used to making the untidy world of science and technology fit our dreams of what science and technology ought to look like. We describe it as though it were all exact and logical and we maintain this model by continual retrospective re-accounting, blaming all mismatches between dream and reality on human error or villainy [2, 7]. To fit a computer into an Area II network of human actors is no more difficult than fitting a slide-rule, a set of log tables, or the activities of scientists and engineers when seen from a distance. We have been doing all this for so long that it comes quite naturally and we never notice it is *we* who are making the pieces fit.

Area IV is much the same except that to make computers work we would have to do much more digitisation and repair. For example, in the case of natural language interactions with computers we would have to learn to make our speech mimeomorphic, and that would mean restructuring our lives. We are unwilling to do this kind of thing and that means computers don't work in the corresponding roles.

That is how the two treatments—the Dreyfus, formal–informal dichotomy scheme, and the dichotomy of actions scheme—come so close to coinciding in terms of their predictions. Nevertheless, under my treatment there are elements of Area IV tasks that can be done by computers and these might grow or diminish as we change our view of how certain tasks ought to be done. What is more, computers will "work" better at Area IV tasks if we are ready to accept ramified behaviors as equivalent to actions; this is a matter of how vigilant or charitable we are ready to be in respect of computers in our midst. If we are very charitable, and ready to do more and more repair work, we may think of computers as accomplishing Area IV tasks. DOCTOR is very cleverly designed to encourage us to accept it in this way. As far as some people are concerned, DOCTOR does passable psychotherapy, and that is certainly an Area IV task.

To conclude, we can see how it is that Dreyfus' predictions are right—or nearly right—in spite of the over-simplification. It is just that in Areas I and II we do not notice how much work we do in embedding computers in our society, whereas in Area IV we are generally unwilling to do the work. For Dreyfus, the combination of professional and sociological errors leads him to think there is a knowledge barrier between domains in which computers can work and those in which they can't, but the areas are not discontinuous—the extent to which computers work within them is a matter of the way *we* act, not the intrinsic nature of domains. This means that Dreyfus does not allow enough scope for computers to creep across the knowledge barrier as clever designers find different ways of substituting for polimorphic actions, as they approximate polimorphic actions with ramified behaviors so that irritating breakdowns become rarer, as they learn to leave it to the user to cope with the inevitable breakdowns, as they find ingenious ways of mechanising more of the little bits and pieces of mimeomorphic action which fill domains like currants in a bun, as they encourage us to be more charitable to their creations, and as they persuade us to do things in different ways so that we can avail ourselves of the economic efficiency of machines.

Part II

7. Individuals and the problem of neural nets

The innovation in the 1992 edition of *What Computers "Still" Can't Do* is the introductory chapter, its second half being about neural nets. Dreyfus appears to have a schizophrenic relationship with neural nets. At the end of the new chapter he returns to the position set out in an earlier article [10]. The crucial insight of the earlier paper is that neural nets learn through stimulus–response [S–R] type training and that this is not equivalent to socialisation. Even the best neural net we have—the embodied human brain—cannot learn human-type situated responses, pattern recognition, and so forth through S–R training and therefore we should not expect it from neural nets. This position is summed up in the very last sentence of the new Introduction: ". . . as improbably as it was that one could

build a device that could capture our humanity in a physical symbol system, it seems at least as unlikely that one could build a device sufficiently like us to act and learn in the world." Actually, most of the preceding discussion in that chapter is about artificial creatures finding their way about in a physical universe and as was explained above in the section on embodiment, this exhibits the philosophical elision. If, however, we take it that succeeding at "acting and learning in the world" could be achieved by learning our language in a "Madeleine-like" way, then Dreyfus can be taken to be talking about the problem of linguistic socialisation.

So far so good, but here and elsewhere Dreyfus recommends neural nets as an improvement on "Good Old Fashioned AI" (GOFAI), because they do not have to represent the world before they can manipulate it and, what may amount to the same thing, because they do not have to be programmed explicitly with rules. There is something wrong with this argument. There are previous generations of intelligent machines that do not have to be programmed with explicit rules, for example rule-inducing expert systems, or record-and-playback robots, that we do not count as breakthroughs in artificial intelligence [3]. Dreyfus does not make clear what is the qualitative difference between these and neural nets.

There is an old saying "my enemy's enemy is my friend". It is almost as though Dreyfus is so concerned with killing GOFAI once and for all that he is ready to embrace GOFAI's enemy, neural nets, in spite of his more general arguments!

I believe, however, that his defence of neural nets is better understood as, again, a consequence of the philosophical elision. If, like Dreyfus, one thinks of the problem of embodiment as being primarily about physical form rather than embedding into a form of life, then the key becomes the architecture of individual devices. For this half of the schizophrenic Dreyfus each new way of making a computer is a new challenge. Digital computers are one thing, analogue computers are another, robots are another, and neural nets are something else again. To see the sort of trouble that this can cause one has only to note that most neural net computers are instantiated on digital machines so that it is far from clear what they are in physical terms. (I have had the benefit of reading a revealing email interchange between Dreyfus and Judea Pearl on this matter.) This side of Dreyfus is continually drawn to considering the architecture of computers and how this might compare with the architecture of the brain. In neural net modellers he seems to find an ally for the unsatisfying models of skill acquisition and the location of knowledge put forward in the book jointly written with his brother [9]. As far as I can see, Dreyfus' philosophical position need not lead in this direction. Of course, an author may go where his interests take him, but I think I now see why Dreyfus' essays into the architecture of computers, computer programs, and the brain are so much less convincing than his brilliant insights into our way of being in the world and their consequences for the project of AI. Different types of computer are better at one thing or another, but this is not a fundamental matter. What is fundamental is the method of embedding in society.

Neural nets are undisputably better at some "intelligent" tasks than other approaches. But perhaps this is because they are easier to "program" with a great

deal of complex information—after all, they "grow" their programs after hours and hours of fast, automated iterative development without humans having to think through every inferential step. Looked at this way, they are a kind of electronically instantiated super programming language. But, for the reason Dreyfus gives in the last sentence of his latest introduction, and for the reasons that I have suggested follow from the idea of a form of life, we ought not to expect them to achieve more than can be achieved by stimulus-response training in the foreseeable future. If I am right about the nature of symbol processing computers—that they can only really do what we do when we are acting as though we had been S–R trained—there is an essential continuity between neural nets and the rest of AI.

8. A simple test for socialisation

All this is quite easy to test if anyone ever comes to believe they have socialised a neural net or other machine. The test does not require that a machine have a body, nor that it be able to do more than produce typescript. The test is even simpler than that devised by Turing. The new test requires an uncharitable judge, an intelligent and literate control who shares the broad cultural background of the judge, and the machine with which the control is to be compared. The judge provides both Control and Machine with copies of a few typed paragraphs (in a clear, machine-readable font), of somewhat mis-spelled and otherwise mucked-about English, which neither has seen before. It is important that the paragraphs are previously unseen for it is easy to devise a program to transliterate an example once it has been thought through. Once presented, Control and Machine have, say, an hour to transliterate the passages into normal English. Machine will have the text presented to its scanner and its output will be a second text. Control will type his/her transliteration into a word processor to be printed out by the same printer as is used by Machine. The judge will then be given the printed texts and will have to work out which has been transliterated by Control and which by Machine. Here is a specimen of the sort of paragraph the judge would present.

Mary: The next thing I want you to do is spell a word that means a religious ceremony.
John: You mean rite. Do you want me to spell it out loud?
Mary: No, I want you to write it.
John: I'm tired. All you ever want me to do is write, write, write.
Mary: That's unfair, I just want you to write, write, write.
John: OK, I'll write, write.
Mary: Write.

The point of this simplified test is that the hard thing for a machine to do in a Turing Test is to demonstrate the skill of repairing typed English conversation—the interactional stuff is mostly icing on the cake [3, 4]. The test is designed to draw on all the culture-bound common-sense needed to navigate the domain of

error correction in printed English. This is the only kind of skill that can be tested through the medium of the typed work but it is quite sufficient, if the test is carefully designed, to enable us to tell the socialized from the unsocialized. (It is worth noting for the combinatorily inclined that a look-up table *exhaustively* listing all corrected passages of about the above length—300 characters—including those for which the most appropriate response would be "I can't correct that", would contain 10^{600} entries, compared to the, roughly, 10^{125} particles in the universe. The number of potentially correctible passages would be very much smaller of course but, I would guess, would still be beyond the bounds of brute strength methods. Note also that the correct response—of which there may be more than one—may vary from place to place and time to time as our linguistic culture changes.)

That we have the ability to repair printed English is not a result of our fixed store of knowledge or a prolonged period of "training". And it is not even that we are corrected when we go wrong at each newly encountered example of a problem. It cannot be this, because the new instances are equally new to any potential trainer and we merely beg the question of how the trainer knows the correct answer. In this kind of case, we cope with newly encountered problems first time and mostly without mistakes. It seems to me that if a machine, neural net or otherwise, could pass a carefully designed version of this little test, all the significant problems of artificial intelligence would have been solved—the rest would be research and development.

9. Final remarks

On my account, Dreyfus' book contains some deep faults and some digressions which detract from the main arguments. Nevertheless, more than 20 years on from its first publication, *What Computers Can't Do* still represents a major achievement and is essential reading for those who want to understand the principle philosophical objections to the idea of artificial intelligence. I have tried to show that in spite of Dreyfus' achievements, there is more work to be done if we are to understand our relationship with machines. Dreyfus explains many of the reasons why machines fail; now we have to understand why and how they succeed. If we understand how much is done by humans in successful human–machine interactions we will see that in the short term, success in the design of intelligent machines can be gained with conceptually simple designs that relieve us of only parts of the job. Identifying those parts is a complicated matter; the answer, difficult enough to extract in a fixed environment, varies from context to context and epoch to epoch. In the longer term, the question of the ability of machines to replicate more and more of our abilities comes down to socialisability. It seems to me that we have not taken even the first step in mechanising socialisibility. But, that this claim can be made without universal derision is, in part, the result of the many years of courageous experimentation that make up the history of artificial intelligence.

References

[1] D. Bloor, Wittgenstein and Mannheim on the sociology of mathematics, *Stud. History Philos. Sci.* **4** (1973) 173–191.

[2] H.M. Collins, *Changing Order: Replication and Induction in Scientific Practice* (Chicago University Press, Chicago, IL, 1985; 2nd ed., with a new Afterword, 1992).

[3] H.M. Collins, *Artificial Experts: Social Knowledge and Intelligent Machines* (MIT Press, Cambridge, MA, 1990).

[4] H.M. Collins, Hubert Dreyfus, forms of life, and a simple test for machine intelligence, *Social Stud. Sci.* **22** (1992) 726–739.

[5] H.M. Collins, The structure of knowledge, *Social Res.* **60** (1993) 95–116.

[6] H.M. Collins and M. Kusch, Two kinds of actions: a phenomenological study, *Philos. Phenomenological Res.* **55** (1995).

[7] H.M. Collins and J.T. Pinch, *The Golem: What Everyone Should Know about Science* (Cambridge University Press, Cambridge, 1993).

[8] H.L. Dreyfus, Response to Collins, *Artificial Experts*, *Social Stud. Sci.* **22** (1992) 717–72.

[9] H.L. Dreyfus and S.E. Dreyfus, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* (Free Press, New York, 1986).

[10] H.L. Dreyfus and S.E. Dreyfus, Making a mind versus modelling the brain: artificial intelligence back at a branchpoint, *Daedalus* **117** (1988) 15–43.

[11] H. Gardinkel, *Studies in Ethnomethodology* (Prentice-Hall, Englewood Cliffs, NJ, 1967).

[12] J. Haugeland, *Artificial Intelligence: The Very Idea* (MIT Press, Cambridge, MA, 1985).

[13] T.S. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago Press, Chicago, IL, 1962).

[14] T.S. Kuhn, The function of measurement in modern physical science, *ISIS* **52** (1961) 162–176.

[15] D. Lenat and E. Feigenbaum, On the thresholds of knowledge, *Artif. Intell.* **47** (1991).

[16] D. MacKenzie, *Inventing Accuracy* (MIT Press, Cambridge, MA, 1991).

[17] D.W. Maurer, *The American Confidence Man* (Thomas, Springfield, IL, 1974).

[18] T.A. Postol, Lessons of the Gulf War experience with Patriot, *International Security* **16** (1991) 119–171.

[19] R.M. Stein, Correspondence: Partiot experience in the Gulf War, *International Security* **17** (1992) 199–225.

[20] L.A. Suchman, *Plans and Situated Action: The Problem of Human–Machine Interaction* (Cambridge University Press, Cambridge, 1987).

[21] L. Wittgenstein, *Philosophical Investigations* (Blackwell, Oxford, 1953).

[22] L. Wittgenstein, *Remarks on the Foundations of Mathematics* (Blackwell, Oxford, 1956).

[23] T. Winograd and F. Flores, *Understanding Computers and Cognition: A New Foundation for Design* (Ablex, Norwood, NJ, 1986).